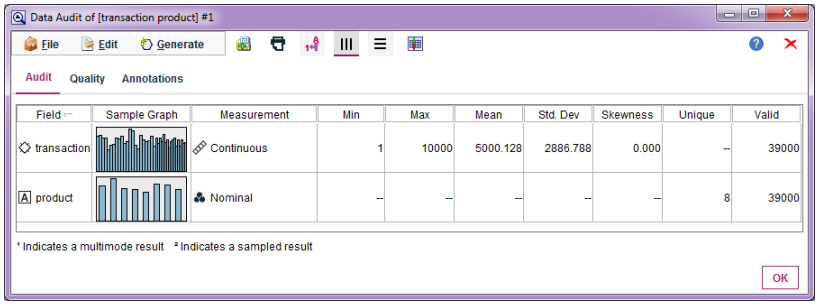
Istraživanje podataka 1 - vežbe 11, 2020.

Primer 1: Primenom pravila pridruživanja proveriti da li postoji zavisnost među podacima u skupu u datoteci *transactions.csv*. Skup *transactions* sadrži podatake u transakcionom obliku. Atributi skupa su:

- *transaction* - identifikator transakcije
- *product* - stavka transakcije

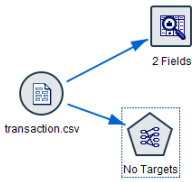
U radnom toku **pravila_pridruzivanja_transactions.str** prvo se učitava skup pomoću čvora *Var. File*. U odeljku *Types* klikom na dugme *Read Values* učitavaju se informacije o vrednostima koje se javljaju u atributima skupa.

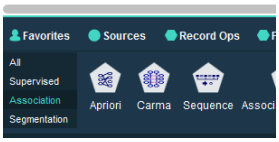
Primenom čvora *Data Audit* prikazuju se osnovne statistike za vrednosti atributa u skupu. Preko kolone *Unique* za atribut *product* dobija se podatak o broju različitih stavki koje se javljaju u skupu (Slika 1).



Slika 1: Prikaz osnovnih statistika za atribute skupa

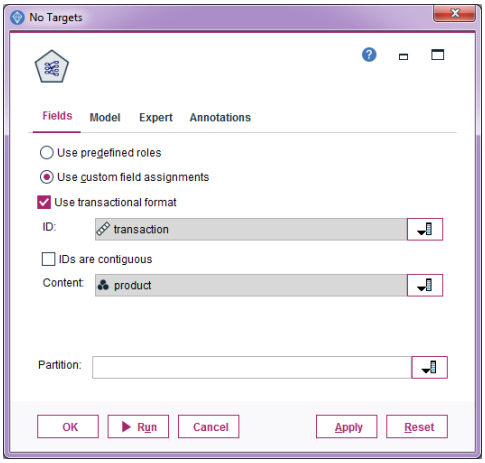
Da bi se primenio algoritm Apriori na skup, čvor sa skupom podataka povezuje se sa čvorom *Apriori* (Slika 2).





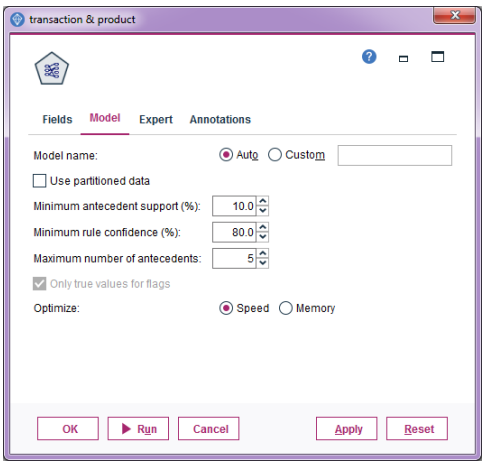
Slika 2: Izbor čvora *Apriori*

Preko opcija dostupnih u čvoru *Apriori*, u odeljku *Fields* navodi se da je skup u transakcionom obliku i postavljaju se podaci koji atribut sadrži identifikator transakcije, a koji atribut sadrži podatke o stavkama. (Slika 3)



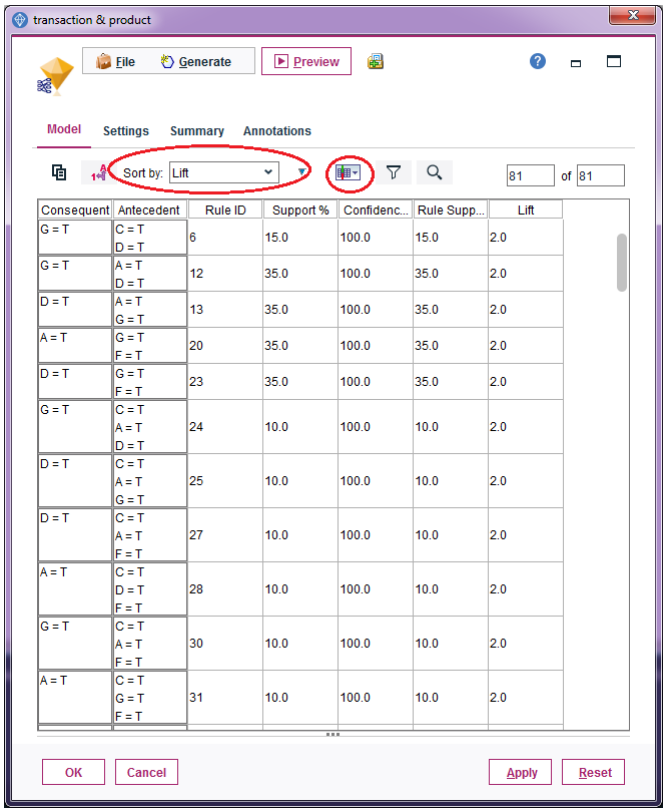
Slika 3: Postavljanje parametara za transakcioni oblik podataka u čvoru *Apriori*

U odeljku *Model*, vrednost za maksimalan broj stavki u telu pravila povećava se na 7 (Slika 4).



Slika 4: Postavljanje vrednosti za maksimalan broj stavki u telu pravila preko čvora *Apriori*

Izborom opcije *Run* pravi se model sa pravilima pridruživanja koji je u radnom toku prikazan čvorom u obliku dijamanta. Duplim klikom na model, u odeljku *Model*, može se videti tabela sa izdvojenim pravilima pridruživanja. Inicijalno se za svako pravilo prikazuju glava pravila, telo pravila, podrška tela i pouzdanost. Klikom na ikonicu *Show/hide criteria menu* biraju se i druge vrednosti koje se mogu prikazati za svako pravilo, npr. podrška pravila, vrednost Lift mere i identifikator pravila. Preko *Sort by* padajućeg menija bira se opcija *Lift* da bi pravila pridruživanja u tabeli bila uređena prema izračunatoj vrednosti za Lift meru (Slika 5).



Consequent	Antecedent	Rule ID	Support %	Confidenc...	Rule Supp...	Lift
G = T	C = T D = T	6	15.0	100.0	15.0	2.0
G = T	A = T D = T	12	35.0	100.0	35.0	2.0
D = T	A = T G = T	13	35.0	100.0	35.0	2.0
A = T	G = T F = T	20	35.0	100.0	35.0	2.0
D = T	G = T F = T	23	35.0	100.0	35.0	2.0
G = T	C = T A = T D = T	24	10.0	100.0	10.0	2.0
D = T	C = T A = T G = T	25	10.0	100.0	10.0	2.0
D = T	C = T A = T F = T	27	10.0	100.0	10.0	2.0
A = T	C = T D = T F = T	28	10.0	100.0	10.0	2.0
G = T	C = T A = T F = T	30	10.0	100.0	10.0	2.0
A = T	C = T G = T F = T	31	10.0	100.0	10.0	2.0

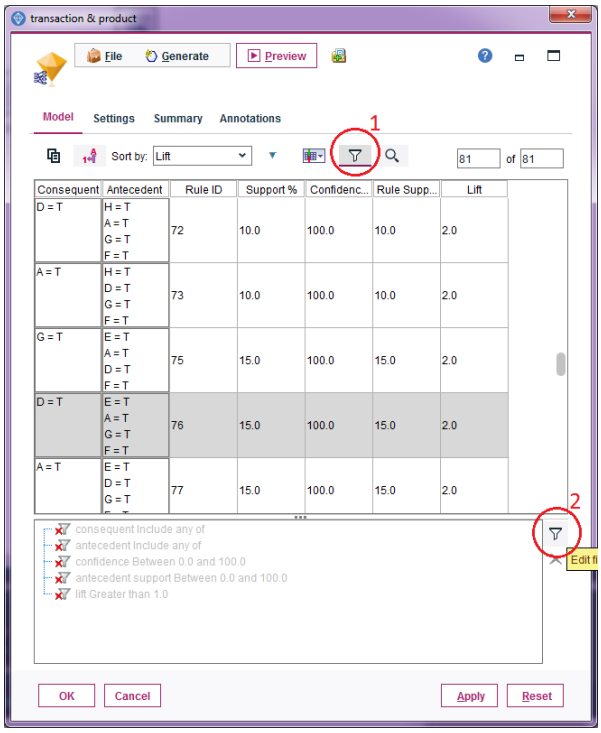
Slika 5: Tabela sa izdvojenim pravilima pridruživanja

Prema Lift meri, sva izdvojena pravila su zanimljiva jer najmanje zanimljivo pravilo ima vrednost 1,455 (>1,1). Najzanimljivija pravila imaju vrednost za Lift meru 2, a među njima su najbolja pravila sa najvećom podrškom pravila (35%) i pouzdanošću (100%):

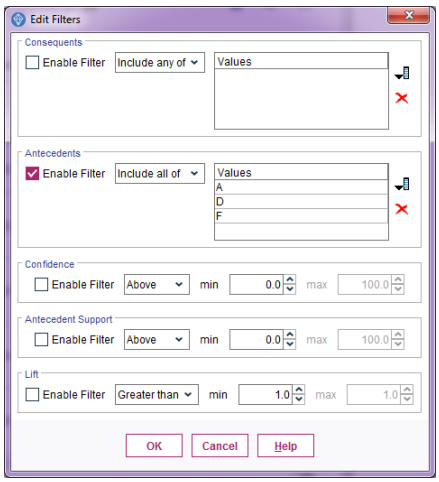
A & D → G
A & G → D
G & F → A
G & F → D

Među najdužim pravilima (sa 4 stavke u telu), najzanimljivija pravila imaju vrednost za Lift meru 2, podršku pravila 15% i pouzdanost 100%:
E & A & D & F → G
E & A & G & F → D
E & D & G & F → A

Ako je potrebno izdvojiti samo pravila koja u telu sadrže stavke A, D i F, primenjuje se filter za pravila pridruživanja. Klikom na ikonicu *Show filters* u donjem delu prozora prikazuju se zadata ograničenja (Slika 6). Klikom na ikonicu *Edit filters* prikazuje se prozor za definisanje želejnih ograničenja za pravila pridruživanja (Slika 7). U tabeli sa pravilima pridruživanja prikazuje se 5 pravila koja zadovoljavaju navedene uslove.

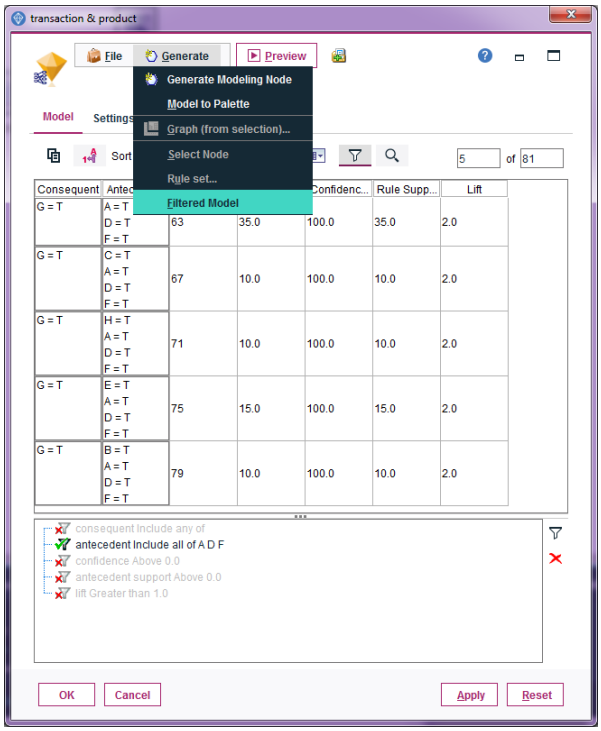


Slika 6: Primena filtera radi izdvajanja pravila pridruživanja sa određenim osobinama



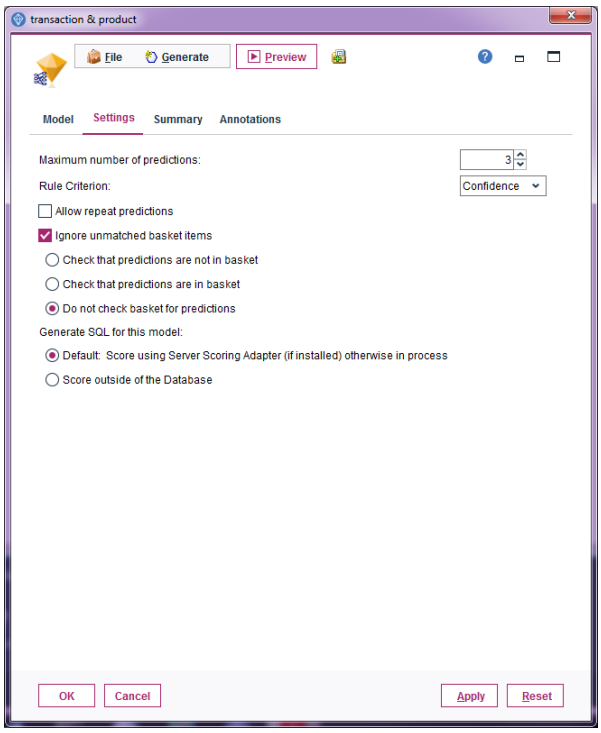
Slika 7: Zadavanje filtera radi izdvajanja pravila pridruživanja sa određenim osobinama

Klikom na dugme *Generate*, a zatim izborom opcije *Filtered Model* u padajućoj list pravi se novi model koji sadrži samo izdvojena pravila. U dijalogu koji se otvara navodi se ime novog modela koji će biti dodat u radni tok (Slika 8).



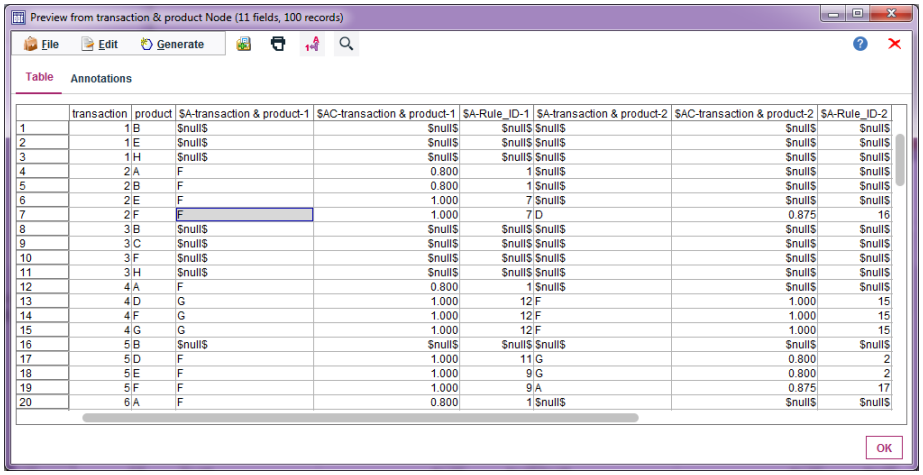
Slika 8: Pravljenje novog modela sa izdvojenim pravilima pridruživanja

U odeljku *Settings* u modelu postavljaju se uslovi za pravila pridruživanja na osnovu kojih se za svaku transakciju pronalazi željeni broj najboljih pravila koja važe u transakciji. Prema postavkama na slici 9 za svaku transakciju se izdvajaju 3 najbolja pravila prema preciznosti; pravila mogu imati istu glavu i glava pravila može, a i ne mora da se javlja u transakciji.



Slika 9: Odeljak *Settings* u modelu sa pravilima pridruživanja

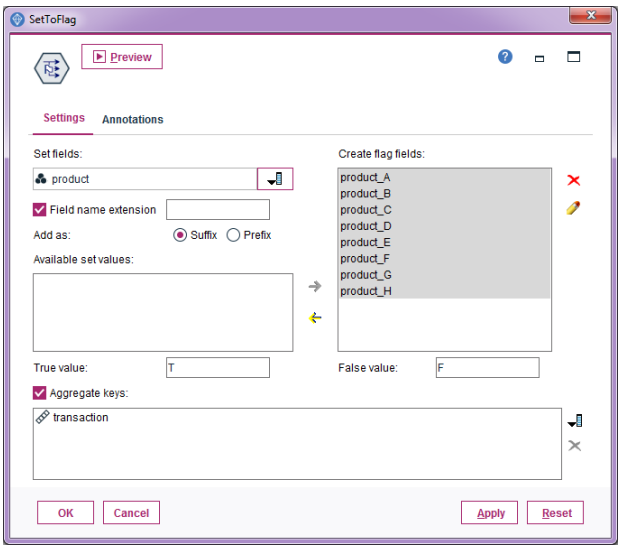
Klikom na dugme *Preview*, prikazuje se tabela sa transakcijama i podaci o najboljim pravilima (Slika 10). Za svako pravilo izdvaja se: glava, pouzdanost i identifikator pravila. Red koji sadrži poslednju stavku transakcije sadrži podatke o najboljim pravilima za celu transakciju (detaljnije objašnjenje o dodeljenim pravilima za ostale redove koji sadrže stavke iste transakcije je u ipVezbe112020Tekst.pdf). Na slici 10, 7. red je poslednji red za transakciju 2 i najbolja pravila prikazana za taj red su najbolja pravila za celu transakciju 2.



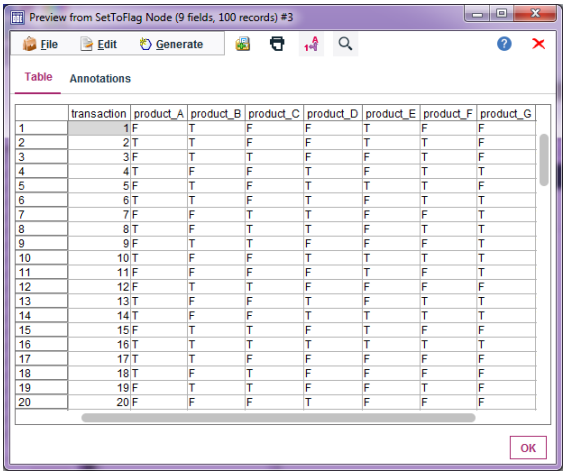
	transaction	product	$A-transaction & product-1	$AC-transaction & product-1	$A-Rule_ID-1	$A-transaction & product-2	$AC-transaction & product-2	$A-Rule_ID-2
1	1	B	$null$	$null$	$null$	$null$	$null$	$null$
2	1	E	$null$	$null$	$null$	$null$	$null$	$null$
3	1	H	$null$	$null$	$null$	$null$	$null$	$null$
4	2	A	F	0.800	1	$null$	$null$	$null$
5	2	B	F	0.800	1	$null$	$null$	$null$
6	2	E	F	1.000	7	$null$	$null$	$null$
7	2	F	F	1.000	7	D	0.875	16
8	3	B	$null$	$null$	$null$	$null$	$null$	$null$
9	3	C	$null$	$null$	$null$	$null$	$null$	$null$
10	3	F	$null$	$null$	$null$	$null$	$null$	$null$
11	3	H	$null$	$null$	$null$	$null$	$null$	$null$
12	4	A	F	0.800	1	$null$	$null$	$null$
13	4	D	G	1.000	12	F	1.000	15
14	4	F	G	1.000	12	F	1.000	15
15	4	G	G	1.000	12	F	1.000	15
16	5	B	$null$	$null$	$null$	$null$	$null$	$null$
17	5	D	F	1.000	11	G	0.800	2
18	5	E	F	1.000	9	G	0.800	2
19	5	F	F	1.000	9	A	0.875	17
20	6	A	F	0.800	1	$null$	$null$	$null$

Slika 10: Prikaz skupa podataka sa pravilima pridruživanja koja zadovoljavaju

Podaci u transakcionom obliku mogu se transformisati u tabelarni oblik korišćenjem čvora *SetToFlag* (Slika 11). Bira se da se izvrši binarizacija kategoričkog atributa *product* i da se koriste sve vrednosti, odnosno stavke, koje se javljaju u atributu, čime se za svaku stavku u skupu podataka pravi jedan atribut. Bitno je postaviti da se vrši grupisanje po transakciji da bi podaci u rezultatu za jednu transakciju bili u jednom redu. Atributi pridruženi stavkama koje se pojavljuju u transakciji imaju vrednost *tačno*, a ostali *netačno*. Ukoliko se ne izvrši grupisanje po transakcijama, jedan red u rezultatu će odgovarati jednoj stavci u transakciji. Deo podataka dobijenih transformacijom je prikazan na slici 12.



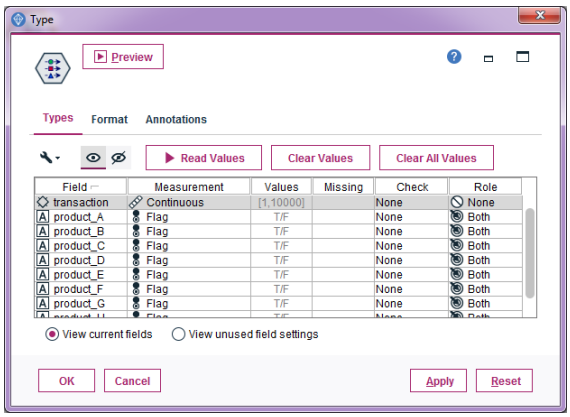
Slika 11: Čvor *SetToFlag* za transformaciju transakcionog oblika skupa u tabelarni



	transaction	product_A	product_B	product_C	product_D	product_E	product_F	product_G
1	1	F	T	F	F	T	F	F
2	2	T	T	F	F	T	T	F
3	3	F	T	T	F	F	T	F
4	4	T	F	F	T	F	T	T
5	5	F	T	F	T	T	T	F
6	6	T	T	F	T	F	T	T
7	7	F	F	T	T	F	F	T
8	8	T	F	T	T	F	T	T
9	9	F	T	T	F	F	F	T
10	10	T	F	F	T	T	T	T
11	11	F	F	F	F	T	F	T
12	12	F	T	T	F	F	F	F
13	13	T	F	F	T	F	T	T
14	14	T	F	F	T	T	T	T
15	15	F	T	T	F	T	F	F
16	16	T	T	T	T	T	T	T
17	17	T	T	F	F	F	F	F
18	18	T	F	T	F	F	F	F
19	19	F	T	T	F	F	T	F
20	20	F	F	F	T	F	F	F

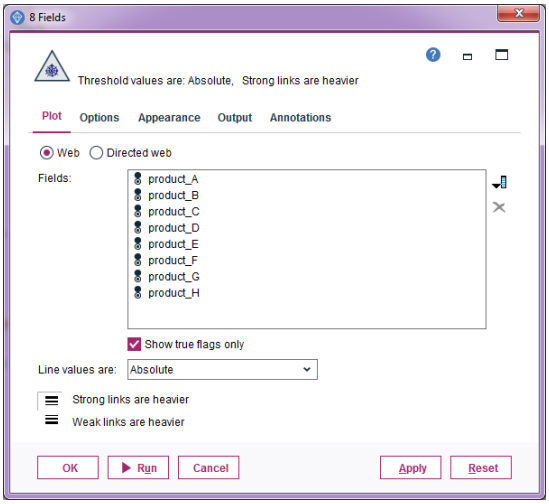
Slika 12: Skup podataka u tabelarnom obliku

Preko čvora *Types* postavljaju se uloge novim atributima. Atribut *transaction* sadrži identifikator transakcije koji nije koristan pri traženju pravila pridruživanja za skup u tabelarnom obliku, te mu se dodeljuje uloga *None*. Atributima koji odgovaraju stavkama u skupu se dodeljuje uloga *Both* da bi svaka stavka iz skupa mogla da se pojavi u telu ili u glavi pravila (Slika 13).



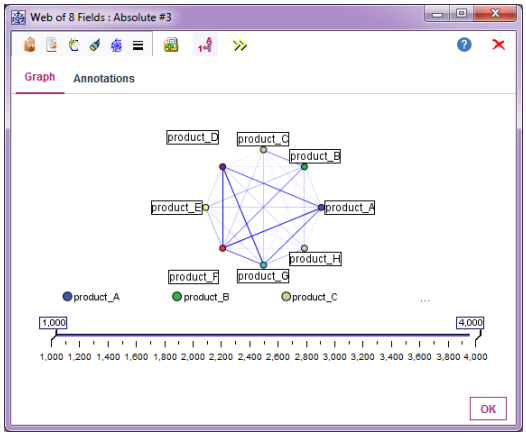
Slika 13: Postavljanje uloga atributima skup podataka u tabelarnom obliku

Za grafički prikaz koliko često se dve stavke javljaju zajedno u transakcijama prikazanim u tabelarnom obliku može da se koristi čvor *Web* iz palete *Graphs*. Preko liste *Fields* u čvoru *Web* se biraju stavke koje će se pojaviti na grafu koji se dobija kao rezultat. Izborom opcije *Show true flags only* na grafu se prikazuju samo veze za stavke koje se javljaju zajedno. Ako se ne izabere ova opcija, za svaku stavku se prave dva čvora; jednom je pridružena vrednost *tačno*, a drugom *netačno*. Nekada je zanimljivo i videti da li postoji jaka veza za pojavljivanje jedne stavke u transakciji i nepojavljivanje druge stavke. Izborom vrednosti *Absolute* za opciju *Line values are* jačina veze između dve stavke se određuje na osnovu broja transakcija u kojima se zajedno javljaju (Slika 14).



Slika 14: Izbor opcija u čvoru *Web*

Klikom na dugme *Run* prikazuje se graf u kome je svaka stavka predstavljena kao jedan čvor. Boja veze između dve stavke je određena brojem transakcija u kojima se te stavke zajedno javljaju. Što je boja veze tamnija stavke se češće pojavljuju zajedno (Slika 15). Prema slici 15 često se zajedno javljaju stavke: D i G, D i A, A i F, D i F.



Slika 15: Prikaz koliko često se dve stavke javljaju zajedno preko grafa

Radi izdvajanja pravila pridruživanja, na čvor *Types* se nadovezuje čvor *Apriori*. U odeljku *Fields* čvora *Apriori* bira se opcija *Use predefined roles* čime se pri traženju pravila pridruživanja koriste ranije dodeljene uloge atributima u čvoru *Types*. U odeljku *Model* i *Expert* opcije se postavljaju kao pri traženju pravila pridruživanja u transakcionom obliku skupa podataka. Tako napravljen model je isti kao model napravljen nad skupom u transakcionom formatu.